A Quadratic Curve Equating Method to Equate the First Three Moments in Equipercentile Equating

Tianyou Wang and Michael J. Kolen

American College Testing

A quadratic curve test equating method for equating different test forms under a random-groups data collection design is proposed. This new method extends the linear equating method by adding a quadratic term to the linear function and equating the first three central moments (mean, standard deviation, and skewness) of the test forms. Procedures for implementing the method and related issues are described and discussed. The quadratic curve method was evaluated using real test data and simulated data in terms of model fit and equating error, and was compared to linear equating, and unsmoothed and smoothed equipercentile equating. It was found that the quadratic curve method fit most of the real test data examined and that when the model fit the population, this method could perform at least as well as, or often even better than, the other equating methods studied. *Index terms: equating, equipercentile equating, linear equating, model-based equating, quadratic curve equating, random-groups equating design, smoothing procedures.*

In standardized testing, multiple test forms are needed because examinees must take the test at different occasions and one test form can be administered only once to ensure test security. Thus, test scores derived from different forms must be equivalent. Efforts can be made in the test construction process to make different forms as nearly equivalent as possible (e.g., forms can be built based on the same table of specifications, or items can be selected to have approximately equal average difficulty level). However, these efforts are often not sufficient to ensure test score equivalence for different forms. Therefore, test equating based on test data is often performed to adjust test scores so that scores on different forms are more nearly equivalent.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 20, No. 1, March 1996, pp 27–43

0146-6216/96/010027-17$2.10

There are several designs for collecting test equating data. One of the designs is the random-groups design, in which different test forms are administered to different but randomly equivalent groups of examinees. Under the random-groups equating design, the examinee groups that take different test forms (for simplicity, say, Form X and Form Y, and Form X scores are equated to the Form Y score scale) are regarded as being sampled from the same population. The differences in score distributions for different test forms are attributed to form differences and sampling variations of the examinee groups. Equating Form X to Form Y involves transforming the X scores so that the transformed Form X scores have the same distribution as the Form Y scores. If an assumption can be made that the population distributions for Form X and Form Y scores have the same shape and only differ in mean and variance, then the linear equating method will be most appropriate. Linear equating (Kolen & Brennan, 1995, p. 30) takes the form

$$l_Y(x) = \sigma_Y\left[\frac{x-\mu_X}{\sigma_X}\right] + \mu_Y, \qquad (1)$$

where

x is the score on Form X,
μ_X and μ_Y are means for Form X and Form Y, respectively,
σ_X and σ_Y are standard deviations (SDs) for Form X and Form Y, respectively, and
$l_Y(x)$ is the equated Form Y score for x.

If no assumptions can be made about the shape of the population score distributions, equipercentile

equating (Braun & Holland, 1982; Kolen & Brennan, 1995, p. 35) is the method of choice. Equipercentile equating for a discrete score distribution is given by

$$e_Y(x) = \frac{p^*(x) - P[Y < u^*(x)]}{P[Y = u^*(x)]} + u^*(x) - .5, \quad (2)$$

where

P means probability,
$p^*(x) = P(X < x) + .5P(X = x)$,
$u^*(x)$ is the smallest integer such that $p^*(x) = P[Y < u^*(x)]$, and
X and Y are random variables for Form X and Form Y scores.

Equipercentile equating based on samples may have large sampling error because for any particular score, the equating relationship is based on local frequencies at that score point. Two types of smoothing techniques have been introduced to reduce random errors: presmoothing and postsmoothing. Presmoothing smooths the score distributions for Form X and Form Y separately prior to equating and equates the smoothed score distributions. Postsmoothing (Kolen, 1984) smooths the equipercentile equating function after unsmoothed equipercentile equating has been implemented.

Studies have been conducted to evaluate the linear, presmoothing, and postsmoothing equating methods (see Cope & Kolen, 1990; Fairbank, 1987; Hanson, 1990; Hanson, Zeng, & Colton, 1991; Kolen, 1984). Results from Hanson et al. (1991) showed that smoothed equating was more accurate than unsmoothed equipercentile and linear methods in terms of mean squared errors. However, linear equating consistently had smaller random error than the pre- and postsmoothing methods, especially when sample sizes were small. This finding resulted because the linear method uses only means and SDs in computing the equating equation and these aggregate statistics typically have small sampling variability. However, a fundamental limitation of linear methods is that if the shape of the distribution of Form X scores is different from that of Form Y scores in the population, the linear equating function could be seriously biased. Although an increase in sample size could reduce standard errors of equating, it will not reduce bias. Angoff (1987) commented that equipercentile equating lacks a theoretical basis whereas linear equating makes strong statistical assumptions that are often violated. He suggested that consideration be given to equating methods that use theoretical models that take into account higher moments. The purpose of this study was to propose a quadratic curve equating method and to compare it to other equating methods. If successful, the quadratic curve method would produce less random error than other pre- or postsmoothing equipercentile methods, and less bias than the linear method.

The Quadratic Curve Equating Method

In selecting a nonlinear equating function, the following aspects were considered:

1. The function should be more flexible than the linear function.
2. The function should preserve beneficial properties of linear equating, such as using statistics with small random errors that are computationally simple.
3. The performance of this method should be comparable to more complicated techniques such as smoothed equipercentile equating in most, if not all, testing situations.

Based on the preceding considerations, a quadratic curve to relate scores on Form X to Form Y was proposed that takes the form

$$q(x) = ax^2 + bx + c. \quad (3)$$

The coefficients a, b, and c are determined such that the equated Form X scores will have the same mean, variance, and skewness as the Form Y scores. The difference between this relationship and the linear equating relationship is that it has one additional squared term and that skewness is taken into account in computing the equating function. The assumption underlying this method is that if population distributions are used and the appropriate quadratic equating relationship is established to equate the first three moments, then all the moments of the equated Form X score distribution will be the same as those of the Form Y score distribution.

The first three central moments (mean, variance,

and skewness) can be equated by equating the first three noncentral moments (moments about the origin) by the definitions of those central moments. For simplicity, the equations that involve noncentral moments are solved to determine coefficients a, b, and c:

$$\mathrm{E}[q(X)] - \mathrm{E}(Y) = 0, \tag{4}$$

$$\mathrm{E}\{[q(X)]^2\} - \mathrm{E}(Y^2) = 0, \tag{5}$$

and

$$\mathrm{E}\{[q(X)]^3\} - \mathrm{E}(Y^3) = 0, \tag{6}$$

where E represents expectation. If $q(X)$ is substituted in these equations,

$$\mathrm{E}[aX^2 + bX + c] - \mathrm{E}(Y) = 0, \tag{7}$$

$$\mathrm{E}[(aX^2 + bX + c)^2] - \mathrm{E}(Y^2) = 0, \tag{8}$$

and

$$\mathrm{E}[(aX^2 + bX + c)^3] - \mathrm{E}(Y^3) = 0. \tag{9}$$

The left-hand sides of these equations are functions of a, b, and c, the first six moments of Form X scores, and the first three moments of Form Y scores. When population distributions are not known, sample moments are used. The multivariable Newton-Raphson iterative method (Press, Flannery, Teukolsky, & Vetterling, 1992, pp. 379–382) can be used to simultaneously solve this set of equations for a, b, and c, iteratively.

Another relatively simple procedure is to find one coefficient at a time. This procedure uses the property that linear transformation does not change the skewness of a score distribution. With skewness of any score distribution Y defined as

$$\mathrm{S}(Y) = \frac{\mathrm{E}[Y - \mathrm{E}(Y)]^3}{\{\mathrm{E}[Y - \mathrm{E}(Y)]^2\}^{3/2}}, \tag{10}$$

this procedure takes the following steps:

Step 1. Using the single-variable Newton-Raphson iterative method (Press et al., 1992, pp. 362–367), find d so that $Z = X + dX^2$ has the same skewness as Y [i.e., S(Z) − S(Y) = 0].

Step 2. Let g be the ratio of the SD of Y to the SD of Z; that is,

$$g = \frac{\mathrm{SD}(Y)}{\mathrm{SD}(Z)}. \tag{11}$$

$gZ = g(X + dX^2)$ will have the same variance and skewness as Y because multiplication by the constant g does not change the skewness of Z.

Step 3. Let $f = \mathrm{E}(Y) - \mathrm{E}[g(X + dX^2)]$. $f + g(X + dX^2)$ has the same mean, variance, and skewness as Y because adding a constant does not change the skewness or variance of a score distribution.

Step 4. The three coefficients in Equation 3 are then computed:

$$a = gd, \tag{12}$$

$$b = g, \tag{13}$$

and

$$c = f. \tag{14}$$

Some Technical Issues

Symmetry. One of the requirements for an equating method is symmetry. That is, the same equating relationship should result whether Form X is equated to Form Y or Form Y is equated to Form X. This quadratic function is clearly not symmetric because different orders of moments are used for Form X and Y scores. Kolen (1984) proposed an average of two equating relations obtained when Form X is equated to Form Y and Form Y is equated to Form X. However, this treatment still does not yield exactly symmetric results.

For the quadratic method, a weighted average of the two equating functions is used. Suppose for a given score x, the equated score obtained from one direction is y_1, that from the other direction is y_2, and the two first derivatives at score point x are d_1 and d_2; then the weighted average is

$$y = w_1 y_1 + (1 - w_1) y_2, \tag{15}$$

where

$$w_1 = \frac{\tan[.5\arctan(d_1) + .5\arctan(d_2)] - d_2}{d_1 - d_2}, \tag{16}$$

if $d_1 \neq d_2$

or

$$w_1 = .5 \quad \text{if } d_1 = d_2. \tag{17}$$

The present authors derived this weight for the linear case. The rationale is to find the line that divides the angle between y_1 and y_2. This weighted average is guaranteed to be symmetric for the linear case. For the quadratic curves in this situation, the curvature can be expected to be very small. Thus, a good approximation to symmetry can be assumed. Note that generally the weights are different at different x scores.

Equating extreme scores. Equating at both ends of the score range is problematic for nearly all equating methods. This is also a problem in the quadratic method. In implementing the postsmoothing method, Kolen (1984) excluded the upper .5% and the lower .5% of the data in computing the postsmoothing function and used two straight lines to link the ends of the equating function to the two unequated end scores. This method was also used here.

Method and Data

Model Fit

Like the linear equating method, the quadratic equating method (QEM) makes an assumption about the true population equating relationship. The assumption underlying the QEM states that the true population equating relationship is quadratic in form. This assumption also implies that after equating using the QEM, all the population central moments of the equated scores of Form X will be the same as those of Form Y. This assumption provides two approaches for examining whether the model fits actual testing data.

One approach is to visually evaluate whether equipercentile equatings (unsmoothed or smoothed) based on real testing data conform to a quadratic form. Even though population score distributions are almost never available in practice, sample distributions (especially large sample distributions) can provide valuable information about model fit. In this study, five different equating methods were applied and plotted for each pair of test forms for visual examination: the unsmoothed equipercentile equating method, the postsmoothing method with smoothing parameter s = .2, the postsmoothing method with smoothing parameter s = .5 (the smoothing parameter s controls the deviation from the unsmoothed function to the smoothed function; see Kolen, 1984, for a detailed description), the QEM, and the linear equating method.

The other approach is to evaluate whether higher central moments (higher than the third moment) after equating also become closer to those of the other test form. Because the fifth or higher central moments of a score distribution have not been defined, only the fourth central moment (kurtosis) was examined here. Kurtosis differences between Form X and Y before and after equating were computed and compared. If the assumption of the QEM is met, the kurtosis of the equated X scores would be expected to be close to that of the Y scores within the limit of sampling error. Under normality, the sampling variance of kurtosis equals $24/N$ where N is sample size (see Kendall & Stuart, 1977, p. 258). The extent to which the model fits the data can be partially assessed by comparing the kurtosis difference after equating and the sampling SD of absolute kurtosis differences.

Datasets

The first two pairs of test score distributions used for examining model fit were the same as the first two pairs used in Hanson et al. (1991). The first pair consisted of two 30-item subsets from a professional licensure exam. The second pair consisted of two 20-item subsets of two forms of the ACT Assessment Program (AAP) Reading subtest. Each of these two datasets had very large Ns (over 38,000 and 82,000 respectively). Thus, the unsmoothed equipercentile equating relationship can be used to approximate the population equating relationship.

Data from an operational equating of the AAP also were used. These data contained score distributions for seven test forms (Form A to Form G) for each of the four tests: English (75 items), Mathematics (60 items), Reading (40 items), and Science (40 items). For each test, seven pairs of distributions were used for equating (Form A was equated to Form B, Form B to Form C ... Form G to Form A). Sample sizes for all equatings are shown in Table 1.

Evaluating Equating Error By Simulation

There are two types of equating errors: random equating error and systematic equating error (see Kolen & Brennan, 1995, for a discussion). Random equating error (indexed by the standard error of equating) is present because sample data are used to estimate population parameters for equating. Systematic equating error (indexed by bias) can be caused by either model misfit of the equating method (i.e., the assumption underlying the method is violated) or by the sampling process (e.g., erroneous sample data or bias embedded in the sampling statistics used to estimate the equating relation). Because the QEM uses aggregate statistics, like the linear method, it was hypothesized that the QEM would have smaller standard errors (SEs) than the unsmoothed equipercentile method or even the smoothed equipercentile methods. It was further hypothesized that the QEM would have smaller bias (in magnitude) than the smoothing methods when the model assumptions were met.

Computer simulation techniques (the parametric bootstrapping method; Efron, 1982) were used to assess the sampling error of the QEM and to compare it to the unsmoothed equipercentile method, the linear method, and the postsmoothing method. With this parametric bootstrapping method, the population score distributions were assumed to be known so that the true equating relationships were also known. Test scores were randomly generated by computer from such populations according to the test score probability distributions, and then various equatings were performed based on the sample distributions and were evaluated against the true equating relationship. More specifically, the simulation in this study took the following steps.

Step 1. A pair of population distributions was defined using either smoothed sample distributions or very large sample distributions. Three types of population distributions were used in this study. The first type was two pairs of observed distributions with very large *N*s. They were the 30-item licensure exam subtests and the 20-item Reading subtests described previously. These observed distributions were taken directly as the population score (probability) distributions.

The second and third types were the results of smoothing AAP score distributions using a loglinear smoothing method (Hanson, 1992; Holland & Thayer, 1987; Kolen, 1991; Livingston, 1993). The second type was intended to represent situations in which the equipercentile equatings with smoothed score distributions were close to the quadratic function. The third type was intended to represent situations in which the equipercentile equatings with smoothed score distributions were not close to the quadratic function. The third type was used also to assess the robustness of the QEM to model violation.

From initial examination of the equating functions from different methods, English Form A and Form B, and Science Form G and Form A were selected to represent the second type; Reading Form A and Form B were selected to represent the third type. Pearson χ^2 statistics for model fit were examined to determine the degree of the loglinear model. The selected degree of the loglinear model is the one for which an increase of one more degree does not bring a significant decrease in the χ^2 value with one degree of freedom. Five pairs of population distributions were used for simulating data.

Step 2. Three different *N*s were used to represent small, medium, and large samples. For the long test (75 items) $N = 500$, 2,000, and 3,000; for short tests of 20, 30, and 40 items, $N = 250$, 500, and 2,000. The sample sizes were different for long and short tests so that frequencies at each score point would be similar. Test scores were computer generated from the population distributions, and equatings with various methods were performed based on the sample score distributions. To generate a test score based on a score probability distribution, a random number from the uniform [0, 1] distribution was generated, and this number was compared to the cumulative score distribution. If the uniform number fell between the cumulative distribution values at score x and $x + 1$, then a score of $x + 1$ was assigned to this simulee. After a pair of test score samples was generated, the equating functions were computed based on five equating methods: the QEM, the linear method, the unsmoothed equipercentile method, the postsmoothing method

with smoothing parameter $s = .2$, and the postsmoothing method with $s = .5$.

Step 3. The second step was repeated $n = 200$ times and evaluative indexes were computed. The true population equating function that was used to compute these indexes was defined as the equipercentile equating function based on the population score distributions.

Evaluative Indexes

The evaluative indexes were bias, SE, and root mean squared error (RMSE). These indexes were evaluated conditionally at all number-correct score points and averaged across the entire score scale. The conditional indexes are defined as follows.

For any score x on Form X, denote $e(x)$ as the true (or population) equated score and $\hat{e}_s(x)$ as the equated score based on sample s with any particular equating method. The mean equated score based on n samples is

$$\bar{\hat{e}}(x) = \frac{1}{n}\sum_{s=1}^{n}\hat{e}_s(x). \tag{18}$$

The estimated bias is

$$\text{Bias} = \bar{\hat{e}}(x) - e(x). \tag{19}$$

The estimated SE is

$$\text{SE} = \left\{\frac{1}{n}\sum_{s=1}^{n}\left[\hat{e}_s(x) - \bar{\hat{e}}(x)\right]^2\right\}^{1/2}. \tag{20}$$

The estimated RMSE is

$$\text{RMSE} = \left\{\frac{1}{n}\sum_{s=1}^{n}\left[\hat{e}_s(x) - e(x)\right]^2\right\}^{1/2}. \tag{21}$$

These conditional indexes then were averaged across the entire score scale using the Form X population score frequencies as the weights. This weighted average was computed as

$$\text{A} = \sum_{x=1}^{k}(\text{index})P(X = x) \tag{22}$$

where k is the number of score points. To compute the average bias, the absolute value of the bias was used because otherwise positive and negative values could cancel. Using the score distribution $P(X = x)$ as the weights is equivalent to using an unweighted average over all the examinees in the population.

Hanson et al. (1991) showed that presmoothing and postsmoothing yielded comparable results in terms of mean squared error. Thus, only postsmoothing was used to represent smoothed equipercentile methods.

Results

Model Fit

The model fit of the QEM was assessed based on the 30 pairs of real test data. For each pair of test data, the new form was equated to the old form using the QEM. The mean, SD, skewness, and kurtosis were computed for both the original (before equating) scores and equated (after equating) scores.

Table 1 contains the before and after equating central moments and Ns for all the test data. (For each of the four AAP tests, Form A was equated to Form B, Form B to Form C ... Form G to Form A. Thus, "Form A after" should be compared to "Form B before," "Form B after" should be compared to "Form C before," and so on.) Ideally, the first three central moments of the new form after equating should be the same as those of the old form; however, because of the adjustment to achieve symmetry, they were not exactly the same, although the differences were quite small, especially for the mean and SD. For example, in Table 1, the AAP English Form A had SD = 12.753 and skewness = −.320 after equating, which were different from but close to the SD and skewness of Form B after equating (12.755 and −.322).

Examination of the kurtosis differences revealed that for 25 out of the 30 equatings, the kurtosis differences were smaller after equating than before equating. The five exceptions were for AAP Reading—Form B before, Form D before, and Form F before—and for AAP Science—Form D before and Form F before. This result suggests that the QEM tends to pull the kurtosis of the two forms closer. The kurtosis difference after equating can also be compared to the kurtosis difference of two randomly drawn samples from the same population distribution. Based on Kendall & Stuart (1977, p. 258), un-

Table 1
Descriptive Statistics for Observed Data Before and After Quadratic Curve Equating

Test and Form	Mean	SD	Skewness	Kurtosis	Kurtosis Difference Before	Kurtosis Difference After	*N*
Licensure Subtest (30 Items)							
New Form Before	18.880	3.680	−.130	2.786			38,765
New Form After	19.157	3.430	−.304	2.934			
Old Form	19.157	3.430	−.308	3.051	.265	.117	38,765
AAP Reading Subtest (20 Items)							
New Form Before	12.300	3.757	−.205	2.391			82,062
New Form After	12.688	3.580	−.278	2.449			
Old Form	12.688	3.580	−.280	2.522	.131	.073	83,693
AAP English (75 Items)							
Form A Before	48.482	13.088	−.089	2.187	.185	.030	2,968
Form A After	51.325	12.753	−.320	2.373			
Form B Before	51.325	12.755	−.322	2.423	.236	.050	2,748
Form B After	48.571	12.207	−.168	2.286			
Form C Before	48.571	12.207	−.167	2.302	.121	.016	2,921
Form C After	51.156	12.206	−.409	2.553			
Form D Before	51.156	12.205	−.406	2.532	.230	.021	2,903
Form D After	50.741	12.204	−.312	2.436			
Form E Before	50.741	12.204	−.313	2.395	.137	.041	2,880
Form E After	51.273	12.770	−.380	2.465			
Form F Before	51.273	12.770	−.381	2.521	.126	.056	2,853
Form F After	50.070	12.876	−.306	2.428			
Form G Before	50.070	12.876	−.308	2.372	.149	.056	2,800
Form G After	48.482	13.091	−.085	2.217			
AAP Mathematics (60 Items)							
Form A Before	28.463	10.569	.481	2.535	.282	.143	2,968
Form A After	30.300	12.198	.256	2.327			
Form B Before	30.301	12.187	.276	2.148	.287	.179	2,748
Form B After	29.758	11.563	.415	2.281			
Form C Before	29.758	11.567	.424	2.463	.315	.182	2,921
Form C After	31.080	12.741	.179	2.298			
Form D Before	31.082	12.722	.208	2.060	.403	.238	2,903
Form D After	28.937	11.448	.296	2.116			
Form E Before	28.937	11.450	.305	2.367	.307	.251	2,880
Form E After	29.819	11.358	.380	2.444			
Form F Before	29.819	11.358	.380	2.424	.057	.020	2,853
Form F After	30.389	11.109	.321	2.375			
Form G Before	30.389	11.108	.324	2.253	.171	.122	2,800
Form G After	28.463	10.566	.474	2.392			
AAP Reading (40 Items)							
Form A Before	24.804	6.584	−.024	2.375	.108	.059	2,968
Form A After	25.350	7.581	−.065	2.386			
Form B Before	25.350	7.581	−.061	2.117	.258	.269	2,748
Form B After	25.669	6.577	−.141	2.158			
Form C Before	25.669	6.578	−.150	2.466	.349	.308	2,921
Form C After	25.837	6.896	−.185	2.492			
Form D Before	25.837	6.896	−.185	2.459	.007	.033	2,903
Form D After	25.314	6.955	−.099	2.408			
Form E Before	25.314	6.954	−.102	2.312	.147	.096	2,880
Form E After	24.731	6.821	.026	2.275			

continued on the next page

Table 1, continued
Descriptive Statistics for Observed Data Before and After Quadratic Curve Equating

Test and Form	Mean	SD	Skewness	Kurtosis	Kurtosis Difference Before	Kurtosis Difference After	*N*
Form F Before	24.731	6.822	.031	2.385	.073	.110	2,853
Form F After	25.452	6.511	−.139	2.458			
Form G Before	25.452	6.512	−.140	2.483	.098	.025	2,800
Form G After	24.804	6.585	−.022	2.434			
AAP Science (40 Items)							
Form A Before	24.153	6.439	−.192	2.553	.148	.042	2,968
Form A After	22.661	7.077	.200	2.543			
Form B Before	22.659	7.064	.170	2.373	.180	.170	2,748
Form B After	22.227	6.964	.231	2.400			
Form C Before	22.330	6.964	.232	2.431	.058	.031	2,921
Form C After	24.122	6.640	−.044	2.415			
Form D Before	24.122	6.642	−.048	2.496	.065	.081	2,903
Form D After	22.965	6.515	.061	2.477			
Form E Before	22.965	6.515	.060	2.463	.033	.014	2,880
Form E After	22.374	6.334	.175	2.495			
Form F Before	22.374	6.334	.173	2.443	.020	.052	2,853
Form F After	22.439	7.073	.110	2.426			
Form G Before	22.439	7.072	.111	2.405	.038	.021	2,800
Form G After	24.153	6.438	−.191	2.511			

der normality the SE of kurtosis of randomly drawn samples of $N = 2{,}900$ scores is .091. The SE for the kurtosis difference for two independent samples of $N = 2{,}900$ is thus approximately .129. Based on the results in Table 1, for the 28 equatings performed on the AAP data, the average absolute kurtosis difference before equating was .163. The average absolute kurtosis difference after equating was .097, which is smaller than the SE of the sample kurtosis difference. 20 of 28 absolute kurtosis differences after equating were smaller than .129. These results suggest that the QEM fit most of the AAP operational equating datasets reasonably well.

Figure 1 shows the equating functions for the two sets of very large sample data. The vertical axis of these plots represents the equating function minus the identity function (i.e., the difference between the equated scores and the original scores). Here, the unsmoothed equipercentile equatings can be used to approximate population equatings due to the large *N*s. Because the unsmoothed functions were already very smooth, it is not surprising that the postsmoothing functions were close to them. The QEM appeared to fit the two population equatings quite well. The maximum biases were within .2 score points. Table 1 also shows that the kurtosis differences were reduced by approximately half after equating in both cases. For example, the kurtosis differences for the licensure subtest were reduced from .265 to .117.

Figures 2a and 2b show five different equating functions for two AAP operational forms. In Figure 2a (English Form A to Form B), the unsmoothed equipercentile equating is close to a quadratic form. Figure 2b (Reading Form A to Form B) is a typical case in which the unsmoothed equipercentile equating is quite different from a quadratic form. Here the unsmoothed equipercentile equating function displays an S shape rather than a quadratic form. [Wang & Kolen (1994) provide a full set of plots of equating functions for the AAP operational data. In most of those plots, the unsmoothed equipercentile equating curve fluctuated around a quadratic curve as in Figure 1 and in Figure 2a.] In both Figures 1 and 2, sometimes the equating curves display a V shape at extreme scores. This was due to the exclusion of the .5% of data at the two ends of the score scale. A straight line was used to connect the two ends of the curves to the end points.

Sampling Error

As described previously, three types of population distributions were used to generate data to study the sampling errors of the equating methods: (1) large-sample data directly taken as population distributions, (2) smoothed distributions in which the quadratic function fit well, and (3) smoothed distributions in which the quadratic function fit poorly. Figure 1 shows the first type of population (true) equating functions, whereas Figure 3 provides the second and third types of population equating functions. Figures 3a (English Form A to B) and 3b (Science Form G to A) represent two pairs of population distributions of the second type, whereas Figure 3c (Reading Form A to B) represents a pair of population distributions of the third type. From these figures, the bias of each of the equating methods can be observed based on the discrepancy of each equating function from the true population equating function (the solid curve). For example, in Figure 3a, at score point 30, the biases of all methods except the linear method are close to 0; at score point 45, all the methods show negative biases and the bias of the QEM is between post-smoothing .2 and post-smoothing .5. The bias of the linear method is very large except at two points (32 and 60).

RMSEs for the equating methods based on four of the five pairs of population distributions are plotted in Figures 4–7. For the first and the third types of populations, only one N was used to exemplify the performance of the QEM (Figures 4 and 7); for the second type of populations, a full set of plots is provided (Figures 5 and 6).

Figure 4 illustrates that, for the licensure test, the QEM performed better than the other methods for much of the score scale for $N = 2{,}000$. In this case, linear equating had a remarkably small RMSE, especially when Ns were small. This result is probably due to the small differences between the forms of the licensure exams.

Figures 5 and 6 show RMSE plots for situations in which the quadratic function apparently fit the population equating relationship well. For these two cases, both the smoothing methods and the QEM gave improved results over the unsmoothed equipercentile methods. The amount of improvement of postsmoothing methods is consistent with the results in Hanson et al. (1991). In these two cases, the QEM tended to perform better than all other methods regardless of N. But the better performance was more consistent along the score scale for small samples than for large samples. In both Figures 5 and 6, the RMSE of the linear method displays large fluctuations. These bumps were apparently due to large bias of the linear method at some score ranges (which can be seen in Figure 3). The score points at which the linear method has a small RMSE correspond to the score points that the linear equating line intercepted with the true population equating curve. Likewise, the RMSE of the QEM also displayed some bumps at extreme score levels, which were also due to relatively large bias in those score ranges.

Figure 7 illustrates the performance of different methods for a situation in which the population equating relationship did not fit a quadratic function. Note that all methods except the unsmoothed equipercentile method fluctuate along the score scale. Comparison to Figure 2 indicates that these fluctuations were due to the fluctuations of bias. Apparently, under this situation, there is no advantage to using the QEM over using the unsmoothed equipercentile method. Interestingly, at $N = 500$ the linear method had the smallest RMSE in the middle score range. The smoothing methods also were not better than the linear method.

Comparison of Figures 4–7 to Figures 1 and 3 shows that the performance of the QEM depended largely on the magnitude of its bias. At score ranges in which the bias of the QEM was small, it typically had a smaller RMSE than the smoothing methods, and vice versa.

In general, these figures also show that the linear method tended to be a viable method when N was very small. When N was very large, postsmoothing with a small smoothing parameter (or in some cases even the unsmoothed equipercentile method) tended to be a good selection.

Table 2 contains the average values of absolute bias (AB), SE, and RMSE weighted by the Form X score population frequencies for all five pairs of population distributions. For the first and second

Figure 1
Equating Functions for the Very Large Samples Datasets

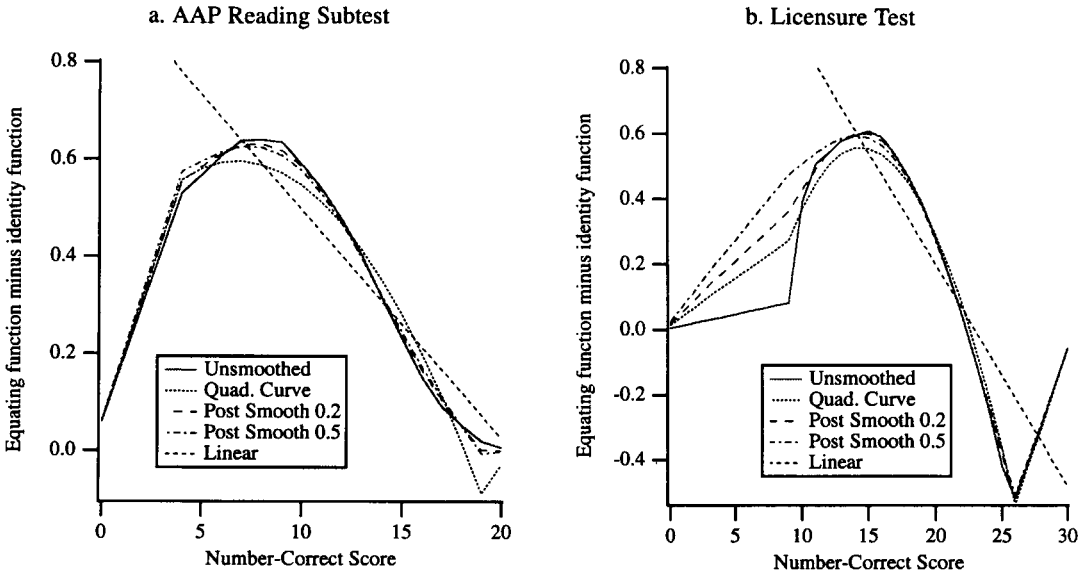
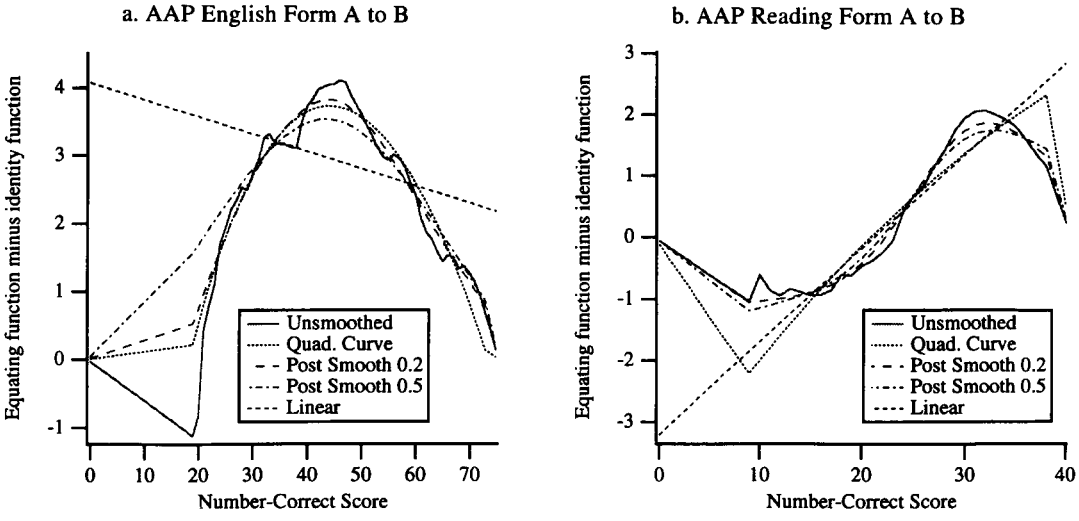


Figure 2
Equating Functions for the AAP Operational Equating Datasets



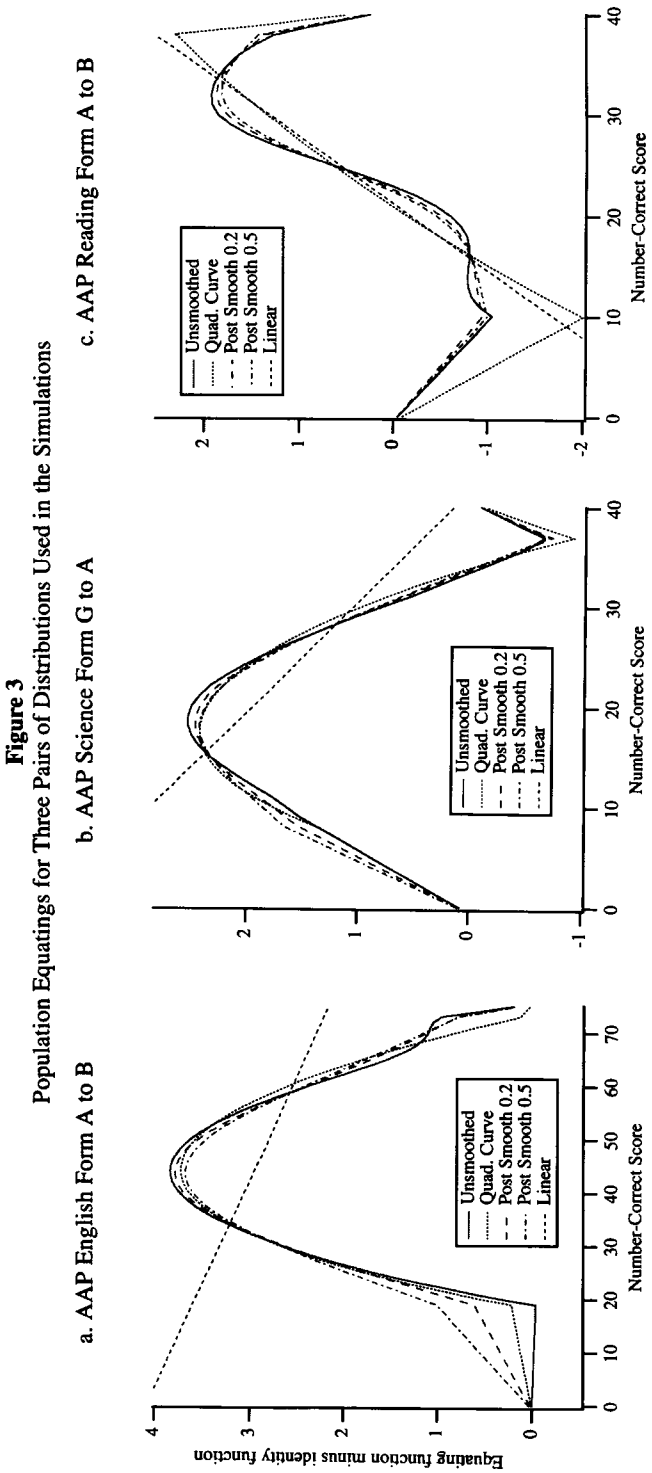
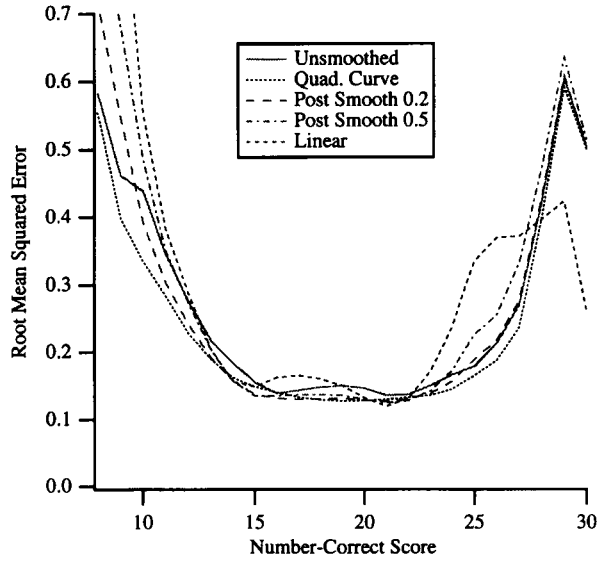


Figure 3
Population Equatings for Three Pairs of Distributions Used in the Simulations

Figure 4
RMSE of Equating Methods for the Licensure Test (N = 2,000)



types of populations (large sample and model fits well), all the SEs were much larger in magnitude than the AB, except for the linear method. Thus, RMSE values were mainly attributed to SEs. Based on RMSE values, Table 2 shows that for the large-sample populations, the QEM had slightly better average performance than the smoothing methods for the licensure subtest with N = 500 or larger, but had slightly poorer average performance for the other cases. For instance, for the licensure subtest with N = 500, the RMSE for QEM was .265, compared to .267 and .269 for the smoothing methods. For the AAP Reading subtest with N = 500, the RMSE for QEM was .266, compared to .259 and .250 for the smoothing methods.

Table 2 also shows that for the "model fits well" populations with small and medium N (N = 500 and N = 2,000 for AAP English; N = 250 and N = 500 for AAP Science), the QEM had smaller RMSE than the smoothing methods. For example, for AAP Science with N = 500, the RMSE for the QEM was .530, compared to .547 and .579 for the smoothing methods. With large sample N (N = 3,000 for AAP English and N = 2,000 for AAP Science), however, the RMSE of the QEM was very similar to that of the smoothing methods.

Table 2 shows that for the "model fits poorly" populations, the QEM had larger average RMSE than the smoothing methods, which was attributed to larger bias. For example, for the AAP Reading test with N = 2,000, the RMSE of QEM was .472, compared to .335 and .374 for the smoothing methods, and this larger RMSE was attributed to the larger AB of QEM (.339 vs. .102 or .192). Postsmoothing with larger smoothing parameters produced a larger bias but a smaller SE than that with smaller smoothing parameters. For example, for the AAP Reading test with N = 2,000, the AB for postsmoothing .2 was .102, compared to .192 for postsmoothing .5, whereas the SE for postsmoothing .2 was .314, compared to .308 for postsmoothing .5. In almost all the cases, linear methods always had smaller SEs.

Discussion and Conclusions

The results of this study show that with a random-groups design, most of the population equating relationships can be approximated by a quadratic function. When a quadratic curve was fit to equate

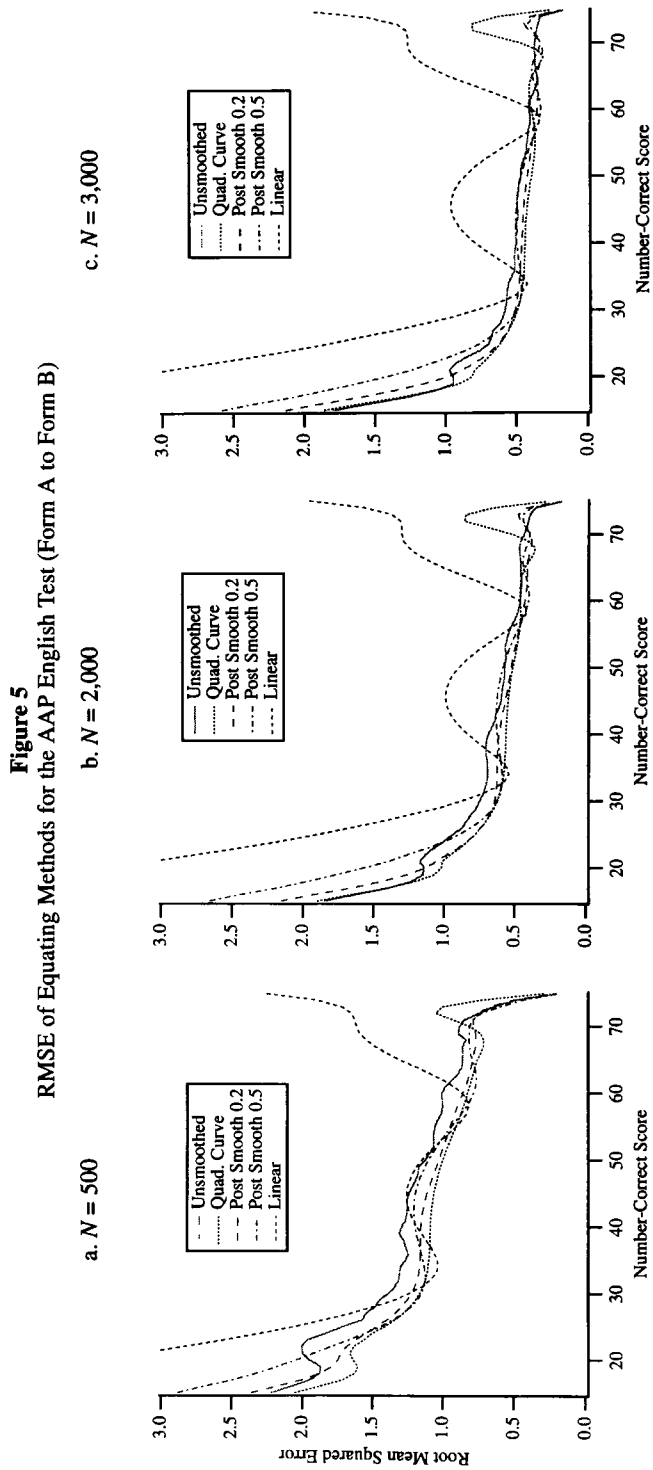


Figure 5
RMSE of Equating Methods for the AAP English Test (Form A to Form B)

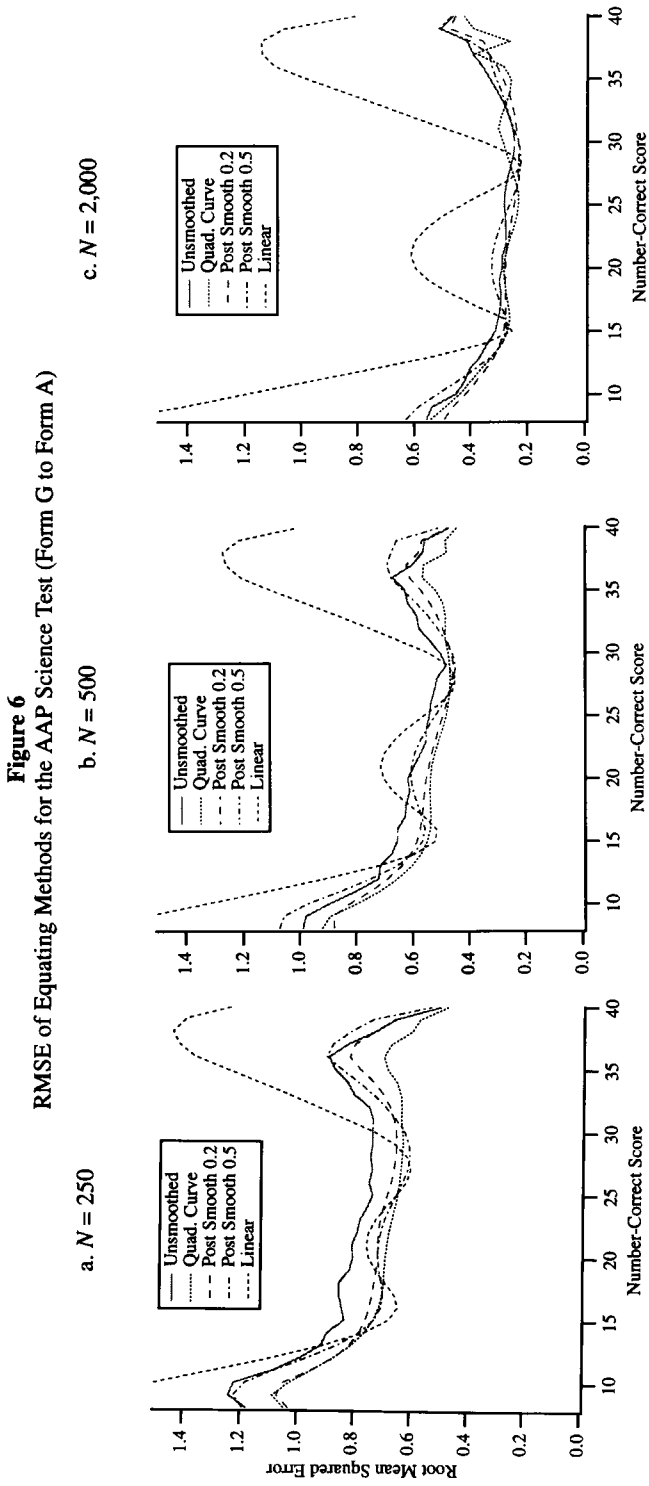
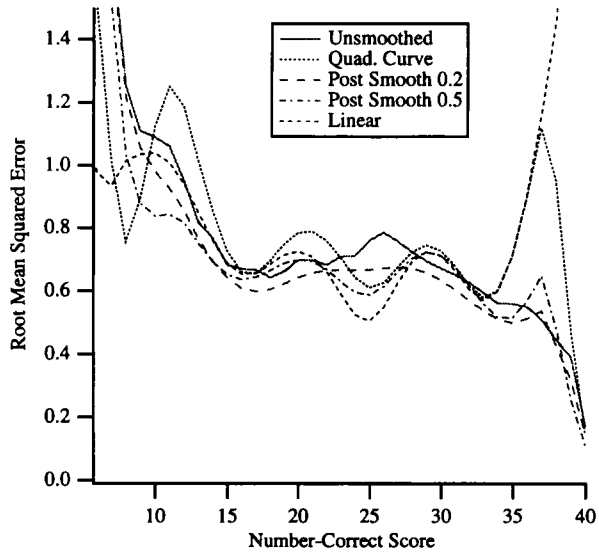


Figure 6
RMSE of Equating Methods for the AAP Science Test (Form G to Form A)

Figure 7
RMSE of Equating Methods for the AAP Reading Test (Form A to Form B) (N = 500)



the first three central moments, the fourth-order central moment (kurtosis) was also equated to some extent in most of the equating data examined. The simulation results show that to the extent the quadratic function fit the population equating function, the quadratic equating method not only showed clear improvement over the linear method and the unsmoothed equipercentile method, but sometimes showed improvement over the postsmoothing methods studied here. The results also show that when N is very small (N = 250 or smaller), the simple linear method provided the best results; when N is very large (2,000 or larger), postsmoothing with a small smoothing parameter (such as .2) is probably a good choice.

In searching for an appropriate polynomial function to model the equating relationship, adding one cubic term to the quadratic function was considered so that kurtosis also could be equated. However, doing so was found to be undesirable for two reasons. First, it makes the computations much more complicated. Second, sample kurtosis has much more random error than skewness. The variance of sample kurtosis for normal distributions is four times that of sample skewness (see Kendall & Stuart, 1977, p. 258). Higher-order polynomial functions might be investigated in the future if these issues can be properly resolved. The quadratic function could be studied as the first step in this direction.

Linear and equipercentile equating both have advantages and limitations. Smoothing methods are aimed at reducing the random error of the equipercentile methods, but they usually involve complicated mathematical manipulation and computer programming. They also often require subjective judgment about model parameters. The quadratic equating method proposed in this paper provides another approach to reducing random error as well as bias. Both the idea and computations are simple, and implementation of the quadratic method does not require subjective judgment.

The results based on the real test data showed that the quadratic method fit most, but not all, of the test data. When the population equating relationship was close to quadratic in form, this method usually displayed smaller random error and bias than other equating methods for Ns in a range of about 250 to 2,000. For N = 250 or smaller, simpler methods such as linear equating show advantages.

Table 2
Average Values of Absolute Bias (AB), Standard Error (SE), and Root Mean Squared Error (RMSE) for Three Population Distribution Types (Large Sample, Model Fits Well, and Model Fits Poorly)

Population Distribution,	$N = 250$			$N = 500$			$N = 2{,}000$		
Test, and Equating Method	AB	SE	RMSE	AB	SE	RMSE	AB	SE	RMSE
Large Sample									
Licensure Subtest									
Unsmoothed	.041	.457	.462	.026	.298	.303	.019	.159	.164
Linear	.113	.388	.410	.117	.242	.276	.111	.132	.180
QEM	.051	.403	.410	.049	.257	.265	.034	.140	.148
Postsmoothed .2	.024	.405	.409	.049	.257	.267	.024	.143	.150
Postsmoothed .5	.069	.387	.399	.088	.247	.269	.059	.140	.158
AAP Reading Subtest									
Unsmoothed	.019	.436	.436	.013	.301	.301	.010	.151	.152
Linear	.062	.365	.373	.063	.249	.259	.068	.124	.145
QEM	.028	.381	.383	.031	.264	.266	.039	.132	.139
Postsmoothed .2	.034	.380	.382	.026	.258	.259	.023	.130	.133
Postsmoothed .5	.051	.361	.366	.047	.244	.250	.046	.122	.133
Model Fits Well									
AAP English Test (Form A to B)[a]									
Unsmoothed	.081	1.146	1.154	.072	.592	.598	.038	.470	.476
Linear	.750	.919	1.230	.685	.471	.904	.733	.370	.849
QEM	.150	.978	1.001	.123	.504	.538	.121	.396	.434
Postsmoothed .2	.102	1.029	1.039	.073	.530	.541	.066	.419	.433
Postsmoothed .5	.334	.984	1.052	.218	.513	.566	.193	.406	.462
AAP Science Test (Form G to A)									
Unsmoothed	.078	.809	.816	.024	.600	.603	.020	.298	.302
Linear	.449	.642	.813	.461	.483	.695	.463	.238	.539
QEM	.101	.680	.694	.101	.516	.530	.103	.254	.281
Postsmoothed .2	.056	.721	.728	.072	.540	.547	.053	.267	.276
Postsmoothed .5	.201	.688	.726	.215	.527	.579	.132	.261	.299
Model Fits Poorly									
AAP Reading Test (Form A to B)									
Unsmoothed	.056	.969	.973	.042	.685	.690	.027	.341	.346
Linear	.351	.787	.883	.355	.564	.689	.359	.283	.476
QEM	.339	.854	.932	.336	.620	.722	.339	.301	.472
Postsmoothed .2	.119	.866	.877	.101	.623	.635	.102	.314	.335
Postsmoothed .5	.258	.787	.839	.244	.587	.647	.192	.308	.374

[a]For the AAP English Test (Form A to B) and the Model Fits Well population distribution, the sample sizes were $N = 500$, $N = 2{,}000$, and $N = 3{,}000$.

For $N = 2{,}000$ or larger, more sophisticated methods such as post-smoothing show advantages. In other cases, the more sophisticated methods displayed smaller random error and bias. Procedures need to be derived to judge whether or not the quadratic method adequately fits the population based on sample data. An examination of the equipercentile equating relationship and the kurtosis difference before and after the quadratic equating might be helpful if this procedure is to be used in practice.

References

Angoff, W. H. (1987). Technical and practical issues in equating: A discussion of four papers. *Applied Psychological Measurement, 11,* 291–300.

Braun, H. I., & Holland, P. W. (1982). Observed score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin

(Eds.), *Test equating* (pp. 9–49). New York: Academic Press.

Cope, R. T., & Kolen, M. J. (1990). *A study of methods for estimating distributions of test scores* (American College Testing Research Rep. 90-5). Iowa City IA: American College Testing.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans.* Philadelphia PA: Society for Industrial and Applied Mathematics.

Fairbank, B. A. (1987). The use of presmoothing and postsmoothing to increase the precision of equipercentile equating. *Applied Psychological Measurement, 11,* 245–262.

Hanson, B. A. (1990). *An investigation of methods for improving estimation of test score distributions* (American College Testing Research Rep. 90-4). Iowa City IA: American College Testing.

Hanson, B. A. (1992). *Description of a program for smoothing univariate test score distributions.* Iowa City IA: American College Testing.

Hanson, B. A., Zeng, L., & Colton, D. (1991, April). *A comparison of presmoothing and postsmoothing methods in equipercentile equating.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (Research Rep. No. 87-31). Princeton NJ: Educational Testing Service.

Kendall, M., & Stuart, A. (1977). *The advanced theory of statistics.* New York: Macmillan.

Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics, 9,* 25–44.

Kolen, M. J. (1991). Smoothing methods for estimating test score distributions. *Journal of Educational Measurement, 28,* 257–282.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer-Verlag.

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement, 30,* 23–39.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1992). *Numerical recipes in C: The art of scientific computing.* Cambridge: Cambridge University Press.

Wang, T., & Kolen, M. J. (1994, April). *A quadratic curve equating method to equate the first three moments in equipercentile equating.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Acknowledgments

The authors thank Ronald T. Cope and Lingjia Zeng for helpful comments on an earlier version of this paper.

Author's Address

Send requests for reprints or further information to Tianyou Wang, ACT, P.O. Box 168, Iowa City IA 52243, U.S.A. E-mail: wang@act.org.